CLAIMS

1. A process for modeling numerical data from a data set comprising:

collecting data for development of a model with a data acquisition module;

processing the data to enhance its exploitability in a data preparation module;

constructing a model by learning on the processed data in a modeling module;

evaluating the fit and robustness of the obtained model in a performance analysis module;

adjusting the model parameters to select the optimal model in an optimization module, wherein the model is generated in the form of a D^{th} order polynomial of the variables used in input of the modeling module, by controlling the trade-off between the learning accuracy and the learning stability with the addition to the covariance matrix of a perturbation during calculation of the model in the form of the product of a scalar λ times a matrix H or in the form of a matrix H dependent on a vector of k parameters $\Lambda = (\lambda_1, \lambda_2, \ldots \lambda_k)$ where the order D of the polynomial and the scalar λ, or the vector of parameters Λ, are determined automatically during model adjustment by the optimization module by integrating an additional data partition step performed by a partition module which consists in constructing two preferably disjoint subsets: a first subset comprising training data used as a learning base for the modeling module and a second subset comprising generalization data destined to adjust the value of these parameters according to a model validity criterion obtained on data that did not participate in the training, and where the matrix H is a positive defined matrix of dimensions equal to the number p of input variables into the modeling module, plus one.

2. The data modeling process according to Claim 1, wherein the matrix H verifies the following conditions: H(i,i) is close to 1 for i between 1 and p, H(p+1,p+1) is close to 0 and H(i,j) is close to 0 for i different from j.

3. The data modeling process according to Claim 1, wherein the matrix H verifies the following conditions: H(i,i) is close to a for i between 1 and p, H(p+1,p+1) is close to b, H(i,j) is close to c for i different from j and a = b + c.

4. The data modeling process according to Claim 3, wherein the matrix H verifies the following additional conditions: a is close to 1-1/p, b is close to 1, c is close to -1/p.

5. The data modeling process according to Claim 1, wherein the matrix H verifies the following condition: H(p+1,p+1) is different from at least one of the terms H(i,i) for i between 1 and p.

6. The data modeling process according to Claim 1, wherein base data partition is performed by an operator using an external software program.

7. The data modeling process according to Claim 1, wherein the data partition module performs a pseudorandom sampling to construct two subsets from the base data.

8. The data modeling process according to Claim 1, wherein the data partition module performs a pseudorandom sampling to construct two subsets from the base data, while keeping the statistical representativeness of the input vectors in the two subsets.

9. The data modeling process according to Claim 1, wherein the base data partition module performs a sequential sampling to construct two subsets from the base data.

10. The data modeling process according to Claim 1, wherein the base data partition module performs a first split of the data into two subsets, with the first subset comprising the training and generalization data and the second subset comprising the test data.

11. The data modeling process according to Claim 1, wherein the base data partition module performs a sampling of the type selecting at least one sample according to a law programmed in advance for generation of the training, generalization and/or test subsets.

12. The data modeling process according to Claim 1, wherein the optimization module selects the pair of parameters (D, λ) or (D, Λ) that minimizes one or the other of the following quantities:

- mean error on the subset of the generalization data;
- weighted mean error on the subset of the generalization data;
- mean quadratic error on the subset of the generalization data;
- weighted mean quadratic error on the subset of the generalization data.

13. The data modeling process according to Claim 1, wherein the data preparation module performs a statistical normalization of columns of data.

14. The data modeling process according to Claim 1, wherein the data preparation

module fills in missing data by one or the other of the following quantities:

- mean of the value on a column (real type data);
- mean of the value on a subset of a column (real type data);
- most frequent value (Boolean or «nominal» type data);
- most frequent value on a subset of a column (Boolean or «nominal» type data);
- selection of a fixed substitution value;
- estimation of the substitution value on the basis of a modeling as a function of other variables.

15. The data modeling process according to Claim 1, wherein the data preparation module performs detection of outlying data according to one or more of the following criteria:

- data outside a range defined by an operator;
- data outside a range calculated by the system;
- Boolean or enumerated data whose number of occurrences is below a given threshold.

16. The data modeling process according to Claim 1, wherein the data preparation module performs a substitution of outliers by one or the other of the following quantities:

- mean of the value on the column (real type data);
- mean of the value on a subset of the column (real type data);
- most frequent value (Boolean or «nominal» type data);
- most frequent value on a subset of the column (Boolean or «nominal» type data);
- selection of a fixed substitution value;
- estimation of the substitution value on the basis of a modeling as a function of other variables.

17. The data modeling process according to Claim 1, wherein the data preparation module performs a monovariable or multivariable polynomial development on all or part of the input data.

18. The data modeling process according to Claim 1, wherein the data preparation module performs a periodic development of the input data.

19. The data modeling process according to Claim 1, wherein the data preparation module performs an explicative development of the input of date type.

20. The data modeling process according to Claim 1, wherein the data preparation module performs a change of coordinates, stemming from a principal components analysis with possible simplification.

21. The data modeling process according to Claim 1, wherein the data preparation module performs one or more temporal shifts before or after all or part of a column containing time variables.

22. The data modeling process according to Claim 1, further comprising exploration of the preparations by a preparation exploration module which uses a description of the possible preparations provided by the user and an exploration strategy based either on a pure performance criterion in training or in generalization, or on a trade-off between these performances and the capacity of the learning process obtained.

23. The data modeling process according to Claim 1, wherein the modeling further comprises model exploitation performed by an exploitation module which provides monovariable or multivariable polynomial formulas descriptive of the phenomenon.

24. The data modeling process according to Claim 18, wherein the modeling further comprises model exploitation performed by an exploitation module which provides periodic formulas descriptive of the phenomenon.

25. The data modeling process according to Claim 19, wherein the modeling further comprises model exploitation performed by an exploitation module which provides descriptive formulas of the phenomenon containing date developments in calendar indicators.

26. The data modeling process according to Claim 18, wherein the periodic development is a trigonometric development.

27. The data modeling process according to Claim 24, wherein periodic formulas descriptive of the phenomenon use trigonometric functions.

28. The data modeling process according to Claim 1, wherein the data preparation module performs one or more of the following actions on «nominal» data to reduce the

number of distinct values:

- calculation of the amount of information brought by each value;

- grouping with each other the values homogeneous in relation to the phenomenon under study;

- creation of a specific value regrouping all of elementary values not providing significant information on the phenomenon.

29. The data modeling process according to Claim 1, wherein the data preparation module regroups missing, outlying or exceptional data into one or more groups to apply a specific processing to them.

30. The data modeling process according to Claim 1, wherein the data preparation module performs encoding of "nominal" type data as tables of Boolean or real variables.

31. The data modeling process according to Claim 1, wherein the data preparation module calculates for each input variable its explicative power in relation to the phenomenon under study.

32. The data modeling process according to Claim 1, wherein the data preparation module uses segmentation algorithms.

33. The data modeling process according to Claim 1, wherein the data preparation module associates with each value of a «nominal» type datum a numerical value representative of the phenomenon under study.

34. The data modeling process according to Claim 1, wherein the data preparation module uses logical rules stemming from knowledge of the phenomena under study to

encode dated data into Boolean values.

35. The data modeling process according to Claim 1, wherein the data preparation module process flows by identifying periodic due dates and applying to them management rules appropriate to each due date.

36. The data modeling process according to Claim 1, wherein training, generalization and forecasting spaces are not disjointed.

37. The data modeling process according to Claim 1, wherein there is defined a relational structure which contains variables, phenomena and models for storing and managing the base data set and the formulas descriptive of the phenomenon.

38. A device for modeling numerical data from a data sample comprising:

means for collecting input data;

means for processing the input data;

means for constructing a model by learning on the processed data;

means for analyzing performances of the obtained model;

means for optimizing the obtained model, wherein the model is generated in the form of a D^{th} order polynomial of the variables used in input of the modeling module, by controlling the trade-off between the learning accuracy and the learning stability with the addition to the covariance matrix of a perturbation during calculation of the model in the form of the product of a scalar λ times a matrix H or in the form of a matrix H dependent on a vector of k parameters $\Lambda = (\lambda_1, \lambda_2, \ldots \lambda_k)$ where the order D of the polynomial and the scalar λ, or the vector of parameters Λ, are determined automatically during model adjustment by the optimization module by integrating additional means for splitting the data so as to construct two preferably disjoint subsets: a first subset comprising training data used as a learning base for the modeling module and a second subset comprising generalization data destined to adjust the value of these parameters according to a model validity criterion obtained on data that did not participate in the training, and where the matrix H is a positive defined matrix of dimensions equal to the number p of input variables into the modeling module, plus one.